BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Perkembangan teknologi komunikasi *mobile* telah meningkatkan ketergantungan masyarakat terhadap layanan pesan singkat (*Short Message Service/SMS*) sebagai media komunikasi yang cepat dan praktis. Namun, di balik kemudahan tersebut, muncul berbagai ancaman keamanan siber, salah satunya adalah *smishing*. *Smishing* merupakan gabungan dari istilah *SMS* dan *phishing*, yaitu teknik penipuan siber yang memanfaatkan pesan teks untuk menipu pengguna agar mengungkapkan informasi sensitif atau mengunduh perangkat lunak berbahaya [1].

Berbeda dengan *phishing* berbasis *email*, *smishing* memanfaatkan sifat *SMS* yang bersifat langsung dan memiliki tingkat keterbacaan tinggi, sehingga lebih sulit dideteksi dan lebih efektif dalam mengecoh korban [2]. Selain itu, rendahnya biaya layanan pesan massal memungkinkan pelaku kejahatan melancarkan serangan dalam skala besar [3]. Kasus pesan palsu yang mengatasnamakan bank, lembaga pemerintah, maupun layanan pengiriman telah menyebabkan kerugian finansial dan kebocoran data dalam skala *global* [4].

Ancaman *smishing* terus berkembang seiring dengan meningkatnya kecanggihan teknik serangan. Penelitian terbaru menunjukkan bahwa pelaku kini memanfaatkan konten berbasis kecerdasan buatan, teknik penyamaran *URL*, serta variasi bahasa untuk menghindari sistem deteksi konvensional [5][6]. Kondisi ini menuntut adanya sistem deteksi yang adaptif dan cerdas, khususnya yang mampu menangani dataset *SMS* berskala besar dalam lingkungan nyata.

Urgensi penanganan *smishing* juga tercermin dari meningkatnya kerugian finansial akibat serangan ini. Data dari *Federal Trade Commission* menunjukkan bahwa kerugian konsumen akibat penipuan berbasis *SMS* mencapai lebih dari 470 juta dolar AS pada tahun 2024, meningkat hampir lima kali lipat sejak tahun 2020 [7]. Selain itu, laporan *Anti-Phishing Working Group* mencatat lebih dari satu juta serangan *phishing* unik pada kuartal pertama tahun 2025, dengan serangan berbasis perangkat *mobile* sebagai kontributor utama [8]. *Smishing* sendiri menyumbang sekitar 37 persen dari seluruh insiden phishing pada perangkat *mobile*, dipicu oleh tingginya tingkat keterbukaan pesan *SMS* dibandingkan *email* [9].

Berbagai pendekatan berbasis *machine learning* telah dikembangkan untuk mendeteksi *smishing* dengan tingkat akurasi tinggi. Namun, sebagian besar penelitian masih memiliki keterbatasan pada ukuran dataset dan kurangnya aspek transparansi model. Oleh karena itu, penelitian ini dianggap penting untuk mengembangkan dan mengevaluasi model deteksi *smishing* yang tidak hanya akurat, tetapi juga dapat dijelaskan *(explainable)*, sehingga layak diterapkan pada sistem keamanan nyata.

Meskipun berbagai penelitian telah mengembangkan model deteksi *smishing* berbasis *machine learning* maupun *deep learning*, masih terdapat sejumlah celah penelitian (*research gap*) yang perlu dikaji lebih lanjut. Salah satu penelitian berjudul *"Enhancing Cybersecurity: Hybrid Deep Learning Approaches to Smishing Attack Detection"* oleh *Mahmud et al.* (14 November 2024) menggunakan pendekatan *hybrid deep learning* berbasis CNN-BiGRU dengan teknik *Word2Vec embedding* serta implementasi *Explainable Artificial Intelligence* (*XAI*) menggunakan *LIME*. Penelitian tersebut menunjukkan performa klasifikasi yang baik, namun masih memiliki keterbatasan pada ukuran dataset, variasi teknik ekstraksi fitur, serta pendekatan interpretabilitas model yang digunakan.[10]

Selain itu, penelitian *"SMS Scam Detection Application Based on Optical Character Recognition for Image Data Using Unsupervised and Deep Semi-Supervised Learning"* oleh *Shinde et al.* (20 September 2024) memanfaatkan kombinasi metode *unsupervised learning* dan *deep semi-supervised learning* dengan fitur *TF-IDF* pada dataset yang relatif terbatas. Penelitian tersebut berfokus pada deteksi *SMS scam* berbasis data citra menggunakan *Optical Character Recognition (OCR)*, namun belum mengeksplorasi berbagai teknik *feature engineering* maupun pendekatan *Explainable Artificial Intelligence* untuk meningkatkan transparansi model.[11]

Berdasarkan perbandingan tersebut, terdapat beberapa *research gap* yang menjadi dasar dilaksanakannya penelitian ini. Pertama, penelitian-penelitian sebelumnya masih menggunakan dataset dengan jumlah yang relatif terbatas, yaitu 24.862 data pada penelitian Mahmud *et al.* dan 7.074 data pada penelitian Shinde *et al.*. Keterbatasan jumlah dan keragaman data tersebut berpotensi memengaruhi kemampuan generalisasi model dalam mengenali berbagai pola serangan smishing yang terus berkembang. Oleh karena itu, penelitian ini menggunakan dataset yang jauh lebih besar, yaitu sebanyak 101.025 data yang berasal dari dua sumber *Mendeley* Data dan satu repositori *GitHub*.

Kedua, dari aspek *feature engineering*, penelitian sebelumnya masih

berfokus pada satu pendekatan representasi teks, seperti *Word2Vec embedding* atau *TF-IDF*. Padahal, efektivitas representasi teks sangat berpengaruh terhadap kemampuan model dalam memahami karakteristik linguistik yang membedakan pesan *smishing* dan pesan normal. Untuk mengatasi keterbatasan tersebut, penelitian ini melakukan eksplorasi berbagai kombinasi teknik ekstraksi fitur dan representasi teks guna memperoleh pendekatan yang paling optimal. Hasil eksperimen menunjukkan bahwa kombinasi *Count Vectorizer* dan *TF-IDF N-gram* menghasilkan performa terbaik.

Ketiga, sebagian besar penelitian terdahulu hanya berfokus pada satu kelompok metode klasifikasi tertentu, baik *deep learning*, *unsupervised learning*, maupun *semi-supervised learning*. Pendekatan tersebut belum memberikan gambaran yang komprehensif mengenai performa relatif dari berbagai jenis algoritma dalam mendeteksi *smishing*. Oleh karena itu, penelitian ini melakukan evaluasi terhadap berbagai pendekatan yang mencakup *machine learning*, *ensemble learning*, dan *deep learning* sehingga dapat diperoleh model yang tidak hanya akurat tetapi juga efisien dari sisi komputasi.

Keempat, aspek interpretabilitas model masih menjadi tantangan dalam penelitian deteksi *smishing*. Penelitian Mahmud *et al.* hanya menggunakan *LIME* sebagai metode *Explainable Artificial Intelligence*, sedangkan penelitian Shinde *et al.* belum mengimplementasikan *XAI* sama sekali. Keterbatasan ini menyebabkan proses pengambilan keputusan model belum dapat dijelaskan secara menyeluruh. Untuk mengatasi hal tersebut, penelitian ini mengombinasikan *SHAP* dan *LIME* sehingga interpretasi model dapat dilakukan baik pada tingkat *global* maupun lokal, serta memberikan pemahaman yang lebih mendalam mengenai faktor-faktor yang memengaruhi hasil klasifikasi.

Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi yang lebih signifikan dalam pengembangan sistem deteksi *smishing* yang akurat, adaptif, dan dapat dijelaskan (*explainable*). Selain meningkatkan performa klasifikasi melalui penggunaan dataset yang lebih besar, eksplorasi fitur yang lebih luas, dan evaluasi berbagai pendekatan model, penelitian ini juga berupaya meningkatkan transparansi sistem melalui implementasi *SHAP* dan *LIME* sehingga lebih siap untuk diterapkan pada lingkungan keamanan siber di dunia nyata.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana perbandingan kinerja model *machine learning, ensemble learning*, dan *deep learning* dalam mendeteksi pesan smishing pada dataset *SMS* berskala besar?
2. Bagaimana implementasi metode *Explainable Artificial Intelligence (XAI)* dalam menjelaskan hasil prediksi model *machine learning, ensemble learning*, dan *deep learning* pada deteksi smishing?

1.3 Batasan Permasalahan

Agar penelitian ini lebih terfokus dan terarah, maka batasan masalah yang ditetapkan adalah sebagai berikut:

1. Data yang digunakan berupa dataset pesan *SMS* berbahasa Inggris yang telah diperoleh sebelumnya.
2. Klasifikasi pesan dibatasi pada dua kelas, yaitu *smishing* dan *non-smishing*.
3. Metode yang digunakan terbatas pada tiga model yang telah ditentukan yaitu *logistic regression, random forest,* dan *BiLSTM*.
4. Evaluasi model dilakukan berdasarkan metrik akurasi, presisi, *recall*, dan *F1-score*.
5. Pendekatan *Explainable Artificial Intelligence* (XAI) yang digunakan terbatas pada metode tertentu untuk menjelaskan hasil prediksi model.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah ditetapkan, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Bagaimana perbandingan kinerja model *machine learning, ensemble learning*, dan *deep learning* dalam mendeteksi pesan smishing pada dataset SMS berskala besar?

2. Bagaimana implementasi metode *Explainable Artificial Intelligence (XAI)* dalam menjelaskan hasil prediksi model *machine learning, ensemble learning*, dan *deep learning* pada deteksi *smishing*?

1.5 Manfaat Penelitian

Manfaat yang diharapkan dari penelitian ini adalah sebagai berikut:

1. Memberikan kontribusi akademik dalam pengembangan sistem deteksi smishing berbasis *machine learning*.
2. Menjadi referensi bagi penelitian selanjutnya yang berkaitan dengan keamanan siber dan phishing berbasis *text*.
3. Membantu praktisi keamanan siber dalam memilih model deteksi smishing yang efektif dan transparan.
4. Meningkatkan kesadaran terhadap pentingnya sistem keamanan *SMS* yang andal dan dapat dipercaya.

1.6 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini disusun untuk memberikan gambaran yang terstruktur mengenai isi setiap bab. Adapun sistematika penulisan penelitian ini adalah sebagai berikut:

- **BAB I PENDAHULUAN**
  Bab ini berisi latar belakang penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.
- **BAB II TINJAUAN PUSTAKA**
  Bab ini memuat landasan teori yang mendukung penelitian, meliputi konsep *smishing*, *phishing*, keamanan siber, *machine learning* untuk klasifikasi teks, teknik *feature engineering*, serta *Explainable Artificial Intelligence* (*XAI*). Selain itu, bab ini juga membahas penelitian terdahulu dan posisi penelitian saat ini.
- **BAB III METODOLOGI PENELITIAN**
  Bab ini menjelaskan tahapan penelitian yang dilakukan, meliputi

pengumpulan dan pemrosesan *dataset*, teknik ekstraksi fitur, pemilihan dan implementasi model *machine learning*, proses pelatihan dan pengujian model, serta metode evaluasi kinerja dan implementasi *XAI*.

- **BAB IV HASIL DAN PEMBAHASAN**
  Bab ini menyajikan hasil eksperimen yang diperoleh dari pengujian berbagai model, analisis perbandingan performa berdasarkan metrik evaluasi, serta pembahasan hasil interpretasi model menggunakan pendekatan *XAI*.
- **BAB V KESIMPULAN DAN SARAN**
  Bab ini berisi kesimpulan yang diperoleh berdasarkan hasil penelitian serta saran untuk pengembangan penelitian selanjutnya.